Bioinformatics Advance Access published November 12, 2005

BIOINFORMATICS

Accurate Anchoring Alignment of Divergent Sequences

Weichun Huang [1,2,3], David M. Umbach [1], Leping Li [1*]

[1]Biostatistics Branch, The National Institute of Environmental Health Sciences, RTP, NC 27709 USA, [2]Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27606 USA, [3]Institute for Genome Sciences & Policy, Duke University Medical Center, Durham, NC 27708 USA

ABSTRACT

Motivation: Obtaining high quality alignments of divergent homologous sequences for cross-species sequence comparison remains a challenge.

Results: We propose a novel pairwise sequence alignment algorithm, ACANA (ACcurate ANchoring Alignment), for aligning biological sequences at both local and global levels. Like many fast heuristic methods, ACANA uses an anchoring strategy. However, unlike others, ACANA uses a Smith-Waterman-like dynamic programming algorithm to *recursively* identify near optimal regions as anchors for a global alignment. Performance evaluations using a simulated benchmark dataset and real promoter sequences suggest that ACANA is accurate and consistent, especially for divergent sequences. Specifically, we use a simulated benchmark dataset to show that ACANA has the highest sensitivity to align constrained functional sites compared to BLASTZ, CHAOS and DIALIGN for local alignment and compared to AVID, ClustalW, DIALIGN, and LAGAN for global alignment. Applied to 6,007 pairs of human-mouse orthologous promoter sequences, ACANA identified the largest number of conserved regions (defined as over 70% identity over 100 bp) compared to AVID, ClustalW, DIALIGN and LAGAN. In addition, the average length of conserved region identified by ACANA was the longest. Thus, we suggest that ACANA is a useful tool for identifying functional elements in cross-species sequence analysis, such as predicting transcription factor binding sites in non-coding DNA.

Availability: ACANA software and test sequence data are publicly available at http://raga.statgen.ncsu.edu/ACANA.

Supplementary information: Supplementary materials are available at *Bioinformatics* online.

Contact: li3@niehs.nih.gov

1 INTRODUCTION

Discovering the function of genes and revealing gene regulation networks are important tasks in decoding genome sequences. Conserved protein domains and functional regulatory sites can provide valuable information for inferring gene function and regulatory controls. Comparative analysis of homologous sequences from related species is an efficient way to reveal such functional domains or regulatory elements (e.g., Xie *et al.*, 2005). With the increasing availability of genome sequences from related species, such cross-species comparative analysis has become more powerful and widely used. The success of a comparative analysis is largely dependent, however, on the accuracy of alignment.

The standard pairwise alignment algorithms can be traced back to the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) for global alignment, and the Smith-Waterman algorithm (Smith and Waterman, 1981) for local alignment. Both the Needleman-Wunsch and the Smith-Waterman algorithms use dynamic programming techniques to find the optimal global and local alignments, respectively. To improve alignment quality, Smith and Waterman (1981) introduced the affine gap cost model for dynamic programming algorithms, which allows one to assign a more flexible penalty for a long insertion/deletion hence possibly makes alignments more biologically meaningful. Gotoh (1982) showed that the affine gap cost model can be implemented with equivalent computational complexity to the constant gap cost model. For short and highly similar sequences, these standard deterministic algorithms work well. Because it is very difficult to assign biologically meaningful gap penalties, these standard algorithms may not be reliable in aligning divergent homologous sequences with long insertions or deletions. A large gap penalty could force mismatched alignments instead of inserting appropriate gap segments, whereas a small gap penalty could result in spurious matching of unrelated regions. Furthermore, the computation time used by either the Needleman-Wunsch or the Smith-Waterman algorithm is proportional to the product of the lengths of two sequences and can increase by a factor of about three when the affine gap cost model is applied. Hence, many heuristic algorithms (Batzoglou *et al.*, 2000; Morgenstern *et al.*, 1998; Morgenstern, 1999; Tatusova and Madden, 1999; Brudno *et al.*, 2003*b*) have been developed to increase alignment speed and/or make alignment more biologically meaningful. BLAST/WU-BLAST (Altschul *et al.*, 1990, 1997), PatternHunter (Ma *et al.*, 2002), BLAT (Kent, 2002), and BLASTZ (Schwartz *et al.*, 2003) are index-based fast local search tools for finding homologous segments. WABA (Kent and Zahler, 2000), MUMmer (Delcher *et al.*, 2002), AVID (Bray *et al.*, 2003), and LAGAN (Brudno *et al.*, 2003*b*) are index-based fast global alignment tools, most of which employ tree-structures to efficiently identify highly identical alignment seeds and use chaining strategies to form anchoring regions for a global alignment. We refer readers to the review by Batzoglou (2005) for more details. These fast tools overcome the problems of insufficient speed and memory and the intolerance of long gaps by the standard dynamic programming algorithm, and thus, they can be used to align genome-size sequences with good accuracy. These tools are widely used and highly effective.

In this paper, we present ACANA (ACcurate ANchoring Alignment), an alternative alignment tool for aligning either DNA or protein sequences. Like many fast heuristic algorithms, ACANA uses the anchoring strategy. However, instead of chaining or extending exactly-matched words as anchoring regions, ACANA uses

*to whom correspondence should be addressed

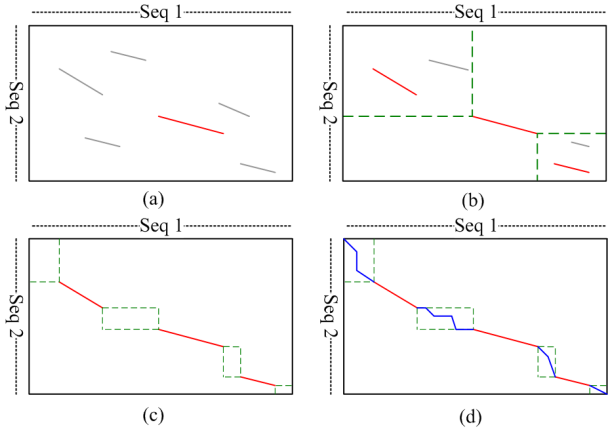

Fig. 1. Illustration of the simplified ACANA algorithm. (a) Compute score of each cell in the matrix by a dynamic programming algorithm, and select the best anchor from local alignments with scores above a certain threshold. (b) Fixing the anchor, recursively select the best anchors in its up-left and down-right regions. (c) All selected anchors are fixed for the global alignment. (d) Finding optimal global alignment for each region between the fixed anchors by Gotoh improved Needleman-Wunsch algorithm, and connecting them with the fixed anchors to generate the global alignment.

the dynamic programming algorithm recursively to identify the near optimal local alignments as anchoring regions. This recursive operation is guaranteed to find near optimal local alignments as it employs the Smith-Waterman algorithm, but it avoids the problem of intolerance of long gaps in sequences.

2 ALGORITHM

ACANA, like many fast alignment tools such as MUMmer, AVID, and LAGAN, uses the anchoring approach for global alignment. However, unlike others, ACANA uses a new strategy for selecting anchoring regions. An anchor-based alignment algorithm has advantages in reducing computation time and/or improving quality of global alignment. ACANA weights local similarity and regional conservation and chooses the best set of anchoring regions from which to construct a global alignment. The simplified ACANA alignment algorithm is illustrated in Figure 1. The four essential parts of ACANA algorithm are: a heuristic dynamic programming algorithm primarily based on the Smith-Waterman algorithm for calculating matrices and tracing local alignment paths using the affine gap cost model; a hash-based algorithm for identifying non-overlapping local alignments in a single pass of calculation of matrices; a method of selecting anchoring regions from local alignments; an algorithm for avoiding unnecessary calculation in recursively searching for the best anchoring regions. Details of these components are described in Implementation. For a pair of sequences, ACANA outputs the non-overlapping local alignments as well as a global alignment, so it is both a local and global alignment tool.

3 IMPLEMENTATION

Calculating Alignment Matrices To find the best local alignment from a pair of sequences A and B of length m and n, respectively, the Gotoh improved Smith-Waterman algorithm (see Supplement) needs to fill three matrices F, G, and H of size $m \times n$ instead of a single matrix by the standard Smith-Waterman algorithm. So it is more computational expensive than the standard Smith-Waterman algorithm. Some improvements have been proposed to increase computational speed (Green, 1993; Trelles *et al.*, 1998; Rognes and Seeberg, 2000), for example, by reducing unnecessary calculations in matrices F and G (Green, 1993). The essential functions of F and G are to store information to decide whether a newly inserted gap extends an existing gap or opens a new gap. That is, F and G are crucial only in locations where insertions or deletions occur. However, significant local alignments generally have few insertions and deletions, so that most elements of F and G are not needed. Our algorithm replaces F and G by a single path-tracing matrix I, and H by a score matrix S. Instead of recording alignment scores, matrix I keeps the path of local alignments. Since the score of a cell in S can only come from three previous cells, two bits of memory is enough for a cell of I to record three possible sources, which can save computation time and space.

ACANA fills S and I by a dynamic programming algorithm with the following recursion relations.

1. IF $i = 0$, set $S_{i,j} = 0$, $I_{i,j} = 0$, where $j = 0...n$
2. IF $1 \leq i \leq m$, calculate $S_{i,j}$ and $I_{i,j}$ by

$$S_{i,j} = \max \begin{cases} c_0 = \max(0, S_{i-1,j-1} + \text{score}(A_i, B_j)) \\ c_1 = S_{i,j-1} + \begin{cases} g_e & \text{if } (I_{i,j-1} = 1) \\ g_o & \text{otherwise} \end{cases} \\ c_2 = S_{i-1,j} + \begin{cases} g_e & \text{if } (I_{i-1,j} = 2) \\ g_o & \text{otherwise} \end{cases} \end{cases}$$

$$I_{i,j} = \begin{cases} 1 & \text{if } (S_{i,j} = c_1) \\ 2 & \text{if } (S_{i,j} = c_2) \\ 0 & \text{otherwise} \end{cases}$$

Where g_o and g_e are gap opening and extension penalties, respectively; $\text{score}(A_i, B_j)$ is the score from a substitution scoring matrix where base A_i is matched with B_j.

ACANA can efficiently track all non-overlapping locally optimal alignments during alignment matrix calculation. The first algorithm for such alignments was introduced by Waterman and Eggert (1987) based on partial recalculation of the alignment matrix H. Barton (1993) extended the algorithm by enabling it to locate all locally optimal alignments in single pass calculation of the matrix H. Barton's (1993) improvement is based on the observation that among all possible paths through each cell $H_{i,j}$, only one gives the optimal alignment. ACANA algorithm is based on the similar observation that there is only one optimal alignment (provided no overlap) from a common start position in the matrix S. When filling the matrix S, ACANA stores the start position of the path passing the current cell using an array of size $\min(m, n) + 1$. In the same time, ACANA employs a hash structure, in which the keys are start positions and the values are stop positions, to dynamically track the stop positions of non-overlapping optimal local alignments. The number of entries in the hash structure is also dynamically updated as new local alignments of scores above a minimum threshold added and those old alignments of scores decaying to zero before reaching a (dynamic) cutoff threshold are removed. In this way, ACANA is able to identify all such locally optimal alignments of scores above a cutoff threshold in a single pass of calculation of alignment matrices. ACANA

uses a new algorithm for tracing the alignment path that is detailed in the Supplement.

Anchor Selection Once a set of significant local alignments in a region has been identified, ACANA selects the best one as the anchor for the region. ACANA's approach is, however, fundamentally different from many existing approaches in which the local alignment with the highest score or the smallest p-value is taken as the best. While existing approaches are valid and efficient, scenarios can arise, especially with two sequences of low to moderate similarity, when the score of a short alignment is larger than that of a much longer alignment that may be biologically more relevant. Thus, it might be a good idea to consider both alignment score and length when choosing an anchor. Herein we propose a new weighted score ω, referred to as a "biological relevance score", for selecting the best local alignment from a set of significant local alignments as an anchor.

$$\omega = \begin{cases} v \log u & \text{if } u \geq 1 \text{ and } v \geq 5 \\ 0 & \text{otherwise} \end{cases}$$

where u and v are the alignment score and length (without counting gaps) of a local alignment, respectively. We believe that length is better than alignment score in assessing biological relevance. So, we choose the length as the main factor of ω, which is then weighted by the logarithm of the local alignment score. Choosing the logarithm of alignment score reflects an idea that for the same amount of increase in score, "biological relevance" should increase more sharply for alignments with small alignment scores than for those with large scores. Use of ω could help ACANA to distinguish significant local alignments from random matched segments and may be important when local alignments are in highly divergent homologous sequences.

ACANA ranks local alignments according to their ω values. If the value of ω for the top ranked local alignment exceeds a certain threshold, ACANA retains, as anchor candidates, all local alignments whose ω values are within 90% percent of that of the top ranked candidate; otherwise no anchor is selected. If there are several anchor candidates, ACANA further calculates a regional weight score for each by $G_a = u_a + \sum_b u_b$, where u_a is the alignment score of the anchor candidate a, and each b is a non-overlapping local alignment that does not intersect with a. G_a is the regional weight score of the anchor candidate a. The candidate with the highest regional weight score is chosen as an anchor for global alignment.

Once the anchor for a region has been selected, ACANA searches all significant local alignments on both sides of the anchor for two new anchors, one for each side. Suppose that the first anchor starts at (a, b) and ends at (c, d) in matrix S, where a and c are the start and stop positions of the anchor in sequence A, and conversely b and d for sequence B. ACANA then searches the up and left corner from rectangular region $(0, 0)$ to (a, b), and right and down corner from (c, d) to (m, n) in matrix S, respectively. This recursive process continues until no anchor can be found.

To find the best local alignment upstream of a fixed anchor, i.e., the region from $(0, 0)$ to (a, b), there is no need to recalculate the scores in the corresponding region in matrix S. However, for the sequence region downstream of the anchor, recalculation becomes necessary. This score recalculation by the dynamic programming algorithm can be costly in the recursive searching process. Fortunately, this step can be avoided for most the cells in region from (c, d) to (m, n), as they do not change during the current iteration of local alignment. Some of these concepts have been previously discussed by Waterman and Eggert (1987). ACANA uses a new algorithm (described in Supplement) to efficiently identify top local alignments within the downstream rectangular region of a fixed anchor by recalculating only some cells in the region. The ACANA matrix recalculation algorithm is actually able to identify top local alignments in any rectangular region of the alignment matrix S.

Construction of Global Alignment After fixing anchoring regions, ACANA uses the Gotoh improved Needleman-Wunsch algorithm to align the remaining sequence segments. These alignments are connected with the anchoring regions to form a global alignment. The default nucleotide substitution scoring matrix of ACANA is based on the alignment-scoring scheme derived by Chiaromonte *et al.* (2002), which is also used by BLASTZ, AVID and LAGAN. The default amino acid substitution scoring matrix is BLOSUM62 from NCBI.

4 RESULTS AND DISCUSSION

Performance Evaluation

To evaluate ACANA's performance, one would ideally apply it to real sequences in which true alignments are known. Although several sets of benchmark protein sequences are available for evaluation of alignment programs (Thompson *et al.*, 1999; Bahr *et al.*, 2001; Lassmann and Sonnhammer, 2002), no good benchmark data from real genomic sequences are currently available. Consequently, we used a benchmark data set of simulated non-coding sequences from Pollard *et al.* (2004) to evaluate the performance of ACANA. In addition, we used data from human-mouse orthologous promoter sequences to assess its performance indirectly.

On Simulated Sequences

In its ability to locally align functionally constrained sites, ACANA compared favorably with DIALIGN (Morgenstern *et al.*, 1996, 1998; Morgenstern, 1999), BLASTZ (Schwartz *et al.*, 2003), and CHAOS (Brudno *et al.*, 2003*a*) on several measures. For example, ACANA had the highest constraint sensitivity among these tools for sequences of intermediate or large divergence differences. Similarly, for global alignment, ACANA appeared to outperform AVID, LAGAN, DIALIGN, and ClustalW (Thompson *et al.*, 1994). Interestingly, the overall sensitivity of ACANA increased as the divergence distance increased whereas the overall sensitivities of the other tools either remained unchanged or decreased. A full exposition of our results on simulated sequences appears in the Supplement.

On Real Sequences

The test data set consists of 6,007 pairs of human-mouse putative orthologous promoter sequences extracted from NCBI GenBank. All known repetitive elements from Repbase database (ver. 8.4) (Jurka, 1998, 2000) were masked in the sequences by Censor (ver. 4.1) (Jurka, 2000) and WU-BLAST (ver. 2.0) (Altschul and Gish, 1996). The list of human-mouse orthologs was from NCBI HomoloGene database. Each promoter sequence is of length 4,500 bp: 3,500 bp upstream and 1,000 bp downstream of the transcription start site as annotated in the GenBank.

Instead of directly measuring alignment accuracy, which is impossible when true alignments are unknown, we assessed global

Table 1. The summary statistics of the average length (*bp*) of conserved regions per pairwise alignment of 6,007 pairs of human-mouse promoter sequences by 5 different alignment tools. A conserved region is defined as more than 70% identity over 100 bp stretch. N is the number of alignments that contain at least one conserved region. SD stands for standard deviation.

	N	Mean	SD	Min	Max
ACANA	4851	907	697	50	4402
AVID	4658	753	614	51	4294
CLUSTALW	4219	893	725	48	4488
DIALIGN	4710	753	621	44	4284
LAGAN	4805	862	677	50	4353

Table 2. The summary statistics of length differences of conserved regions detected by 5 different alignment tools. Here the N is the number of pairs of orthologous genes, from which both alignment tools can find conserved regions. The p-value were computed from a paired t test. SD stands for standard deviation.

Difference	N	Mean	SD	$\Pr > \|t\|$
ACANA - AVID	4641	185	168	$< .0001$
ACANA - CLUSTALW	4190	67	213	$< .0001$
ACANA - DIALIGN	4699	177	167	$< .0001$
ACANA - LAGAN	4788	52	92	$< .0001$

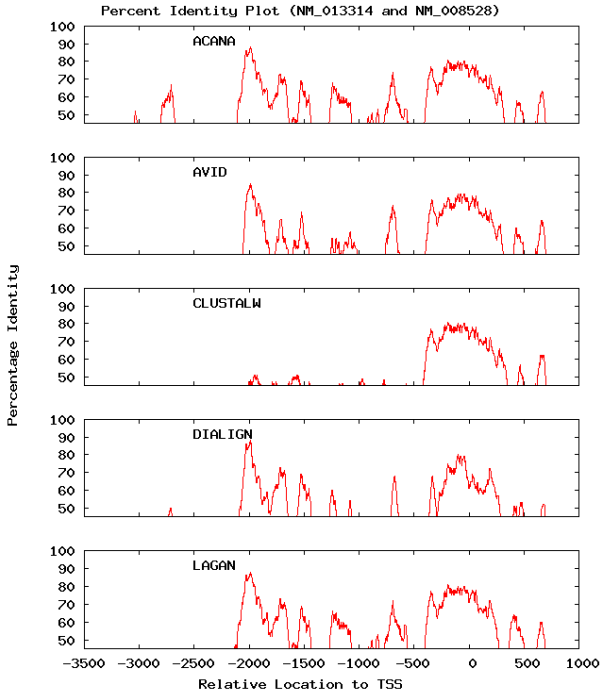

Fig. 2. Percent Identity Plots of global alignments The promoter sequences are of human (NM_013314) and mouse (NM_008528) orthologous genes encoding B-cell linker protein. The plots show that only ACANA is able to detect a conserved region between positions [-3000,-2500] relative to the transcription start site of the human gene.

alignment quality by the relative length of all conserved regions aligned by different tools. From a biological point of view, an accurate global alignment should correctly align evolutionarily related regions, including the syntenically conserved regions. Therefore, the relative length of syntenically conserved regions aligned can be used as an indirect measure for assessing quality of global alignments by different tools.

For global alignment, we compared ACANA with AVID, ClustalW, LAGAN and DIALIGN. For DIALIGN, we used the improved version — DIALIGN-2 (Morgenstern, 1999). First, each tool, with its default parameter settings, was used to align each pair of orthologous sequences. Second, for each pairwise global alignment, we used VISTA (Mayor *et al.*, 2000) to extract conserved regions (see Supplement). Although methods used to define conserved regions are somewhat arbitrary, one of the most frequently used is based on percentage identity over a region of fixed length (Fickett and Wasserman, 2000; Loots *et al.*, 2000). VISTA employs this method with a default cutoff value of 70% identity over 100 *bp*, a value commonly used for human and rodent species.

ACANA not only finds the largest number of orthologous pairs of sequences containing at least one conserved region but also the longest conserved region on average compared to the other two tools (Tables 1 and 2). To see the differences, we randomly picked 100 orthologous pairs of sequences from the data set. For each pair, we manually examined the three alignments by their Percent Identity Plots (PIP). In all cases, the PIP plots of alignments from the three algorithms are similar for sequence regions with high similarity, but may be different for regions with only moderate similarity. An example is given in Figure 2.

Summary and Future Work

A challenge in comparative sequence analysis is to obtain high quality sequence alignments while minimizing computational time. In the past two decades, significant progress has been made. The most important achievement is the dramatic reduction in computation time by heuristic algorithms coupled with faster computers, which makes it possible to align genome-size sequences. Despite this progress, many challenges remain, notably the quality of alignment. Except when applied to the smallest and simplest sequences, almost no two current alignment algorithms regularly give the same alignment. It is very difficult, if not impossible, to reflect accurately evolutionary events such as point mutation, insertion, deletion, duplication, rearrangement, etc. in a scoring function for alignment. Nonetheless, the recent advances in alignment methodologies have made a great impact on modern biological research.

The introduction of anchoring has made genome-wide alignment feasible and fairly accurate. While the index or word-chaining based approaches are very efficient, these heuristic approaches are not guaranteed to find the near optimal local alignments as anchors, especially for divergent sequences. The ACANA algorithm uses the Smith-Waterman-like dynamic programming algorithm for local alignment, enabling it to identify the near optimal local alignments.

Furthermore, ACANA uses a new strategy to select an anchoring region from a set of significant local alignments.

Performance evaluations suggest that ACANA is an accurate and consistent alignment tool for both local and global alignments. Using a set of simulated benchmark dataset, we found that ACANA has the highest constrained sensitivity in correctly aligning the known constrained functional sites embedded in the sequences compared to BLASTZ, CHAOS, and DIALIGN for local alignment and AVID, ClustalW, DIALIGN, and LAGAN for global alignment. ACANA performs best for sequences of moderate to large divergent distances. When tested on a set of paired putative human/mouse orthologous promoter sequences, ACANA found the largest number of orthologs that contained at least one conserved region (over 70% identity over 100 *bp*) compared to AVID, ClustalW, DIALIGN and LAGAN. In addition, the average length of the conserved regions identified by ACANA was the longest. We believe that ACANA shows some improvement over existing tools for aligning divergent sequences. We attribute the potential improvements partially to ACANA's recursive anchoring selection strategy.

We would like to point out that the current version of ACANA is not capable of dealing with inversions in sequences. We think that such capability can be easily incorporated by aligning sequence segments in both directions in the recursive anchoring selection step. Such work is in progress. Lastly, ACANA may be combined with other faster local alignment tools such as CHAOS when significant improvement in speed is needed to align genome-size sequences.

In conclusion, we believe that ACANA is a novel and accurate alignment algorithm. Its new recursive anchoring selection strategy may represent an improvement over existing methods. ACANA's ability to align conserved functional sites and its robustness to large insertions/deletions make it particularly useful in comparative genomic analysis of promoter sequences for functional element discovery.

ACKNOWLEDGMENTS

We thank Bruce Weir and Jeffrey Thorne for critically reading the manuscript, and Clarice Weinberg and Stephen Heber for helpful comments. We would also like to thank the reviewers for their valuable suggestions to improve the presentation of this paper. This research was supported by Intramural Research Programs of the NIH, National Institute of Environmental Health Sciences.

REFERENCES

Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.,* **266**, 460–480.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.,* **215**, 403–410.

Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.,* **25**, 3389–3402.

Bahr,A., Thompson,J.D., Thierry,J.C. and Poch,O. (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res.,* **29**, 323–326.

Barton,G.J. (1993) An efficient algorithm to locate all locally optimal alignments between two sequences allowing for gaps. *Comput Appl. Biosci.,* **9**, 729–734.

Batzoglou,S. (2005) The many faces of sequence alignment. *Brief Bioinform,* **6**, 6–22.

Batzoglou,S., Pachter,L., Mesirov,J.P., Berger,B. and Lander,E.S. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.,* **10**, 950–958.

Bray,N., Dubchak,I. and Pachter,L. (2003) AVID: A global alignment program. *Genome Res.,* **13**, 97–102.

Brudno,M., Chapman,M., Göttgens,B., Batzoglou,S. and Morgenstern,B. (2003*a*) Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics,* **4**, 66.

Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., Green,E.D., Sidow,A. and Batzoglou,S. (2003*b*) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.,* **13**, 721–731.

Chiaromonte,F., Yap,V.B. and Miller,W. (2002) Scoring pairwise genomic sequence alignments. *Pac. Symp. Biocomput.,* **7**, 115–126.

Delcher,A.L., Phillippy,A., Carlton,J. and Salzberg,S.L. (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.,* **30**, 2478–2483.

Fickett,J.W. and Wasserman,W.W. (2000) Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.,* **11**, 19–24.

Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.,* **162**, 705–708.

Green,P. (1993). http://www.genome.washington.edu/uwgc/analysistools/swat.cfm.

Jurka,J. (1998) Repeats in genomic DNA: mining and meaning. *Curr Opin Struct Biol,* **8**, 333–337.

Jurka,J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.,* **16**, 418–420.

Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.,* **12**, 656–664.

Kent,W.J. and Zahler,A.M. (2000) Conservation, regulation, synteny, and introns in a large-scale c. briggsae-c. elegans genomic alignment. *Genome Res.,* **10**, 1115–1125.

Lassmann,T. and Sonnhammer,E.L.L. (2002) Quality assessment of multiple alignment programs. *FEBS Lett.,* **529**, 126–130.

Loots,G.G., Locksley,R.M., Blankespoor,C.M., Wang,Z.E., Miller,W., Rubin,E.M. and Frazer,K.A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science,* **288**, 136–140.

Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics,* **18**, 440–445.

Mayor,C., Brudno,M., Schwartz,J.R., Poliakov,A., Rubin,E.M., Frazer,K.A., Pachter,L.S. and Dubchak,I. (2000) VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics,* **16**, 1046–1047.

Morgenstern,B. (1999) Dialign 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics,* **15**, 211–218.

Morgenstern,B., Dress,A. and Werner,T. (1996) Multiple dna and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci. USA,* **93**, 12098–12103.

Morgenstern,B., Frech,K., Dress,A. and Werner,T. (1998) Dialign: finding local similarities by multiple sequence alignment. *Bioinformatics,* **14**, 290–294.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.,* **48**, 443–453.

Pollard,D.A., Bergman,C.M., Stoye,J., Celniker,S.E. and Eisen,M.B. (2004) Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics,* **5**, 6.

Rognes,T. and Seeberg,E. (2000) Six-fold speed-up of Smith-Waterman sequence database searches using parallel processing on common microprocessors. *Bioinformatics,* **16**, 699–706.

Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.,* **13**, 103–107.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.,* **147**, 195–197.

Tatusova,T.A. and Madden,T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.,* **174**, 247–250.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.,* **22**, 4673–4680.

Thompson,J.D., Plewniak,F. and Poch,O. (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics,* **15**, 87–88.

Trelles,O., Andrade,M.A., Valencia,A., Zapata,E.L. and Carazo,J.M. (1998) Computational space reduction and parallelization of a new clustering approach for large groups of sequences. *Bioinformatics,* **14**, 439–451.

Waterman,M.S. and Eggert,M. (1987) A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J. Mol. Biol.,* **197**, 723–728.

Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature,* **434**, 338–345.